# Assignment 2

BigData Week3

2025. 3.20

Eunhui Kim (김은희)
ehkim@kisti.re.kr

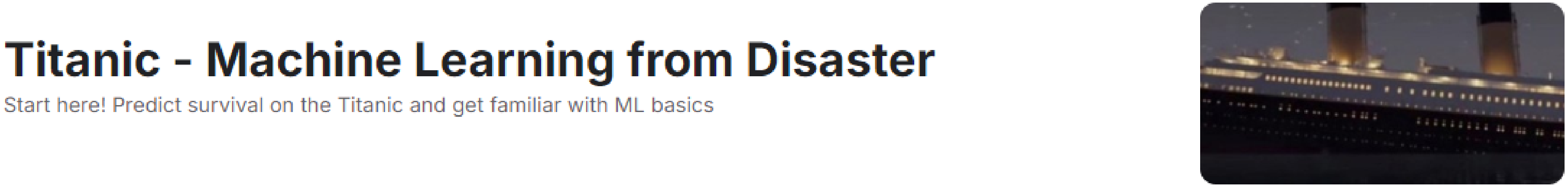
# Assignment 2

- Submission due : March. 27 th, 23:55
- What to submit : Notebook file (.ipynb)
  - Colab : [File]-[Download]-[Download .ipynb]
  - Kaggle : [File]-[Download Notebook]

**IMPORTANT**

- Be sure to download the dataset from Assignment-2
  - The file name is “titanic_rev.csv”
  - This is a modified data different from Assignment-1
- Make sure your results are the same as the output presented on this slide
  - Problem 5-7 : depending on whether you accumulate the result into a single DataFrame, the result may differ

# Assignment 2



## Titanic - Machine Learning from Disaster

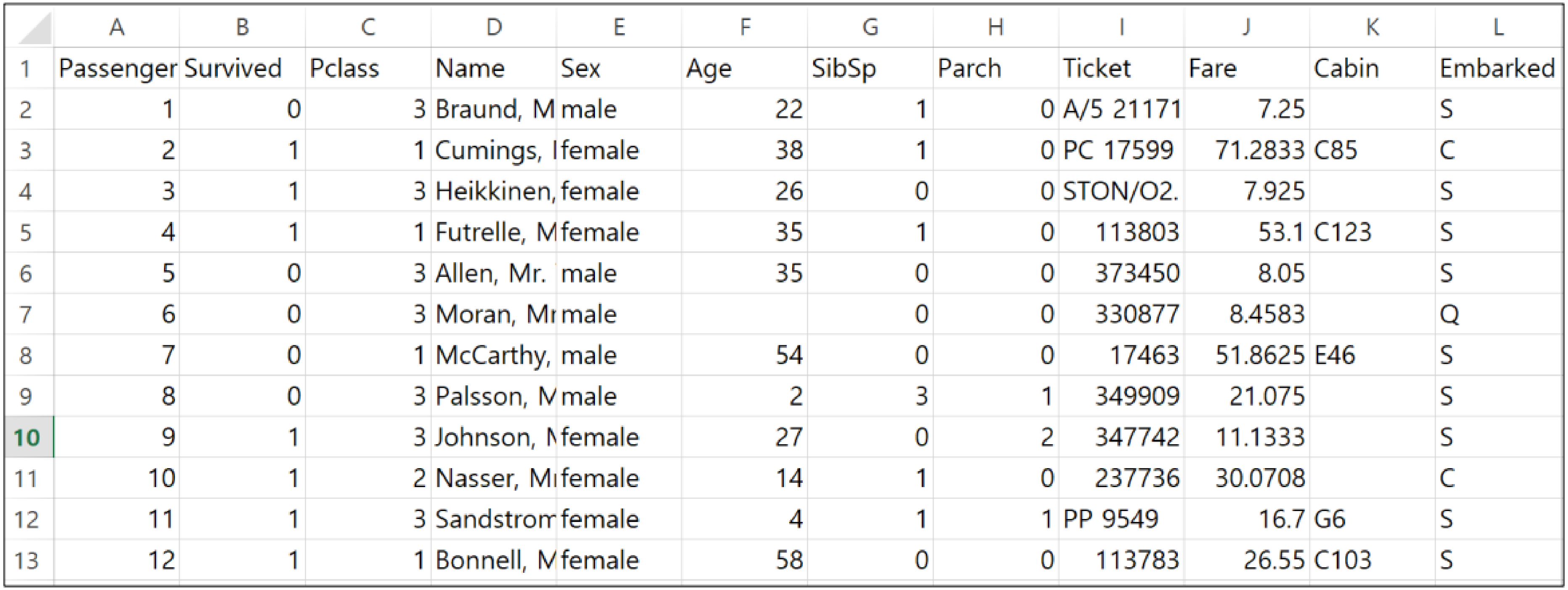
Start here! Predict survival on the Titanic and get familiar with ML basics

# Assignment 2

- Titanic dataset includes:
  - Passenger ID
  - Passenger Class (1st, 2nd, or 3rd class)
  - Name
  - Sex
  - Age
  - Sibling/Spouse Aboard (SibSp)
  - Parent/Child Aboard (Parch)
  - Ticket Number
  - Fare
  - Cabin Number
  - Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
  - Whether the passenger survived (1 for survived, 0 for did not survive)

# Assignment 2

- Titanic dataset includes:



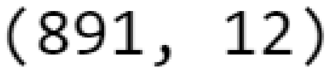
|  | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Passenger | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
| 2 | 1 | 0 | 3 | Braund, M | male | 22 | 1 | 0 | A/5 21171 | 7.25 |  | S |
| 3 | 2 | 1 | 1 | Cumings, I | female | 38 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 4 | 3 | 1 | 3 | Heikkinen, | female | 26 | 0 | 0 | STON/O2. | 7.925 |  | S |
| 5 | 4 | 1 | 1 | Futrelle, M | female | 35 | 1 | 0 | 113803 | 53.1 | C123 | S |
| 6 | 5 | 0 | 3 | Allen, Mr. | male | 35 | 0 | 0 | 373450 | 8.05 |  | S |
| 7 | 6 | 0 | 3 | Moran, Mr | male |  | 0 | 0 | 330877 | 8.4583 |  | Q |
| 8 | 7 | 0 | 1 | McCarthy, | male | 54 | 0 | 0 | 17463 | 51.8625 | E46 | S |
| 9 | 8 | 0 | 3 | Palsson, M | male | 2 | 3 | 1 | 349909 | 21.075 |  | S |
| 10 | 9 | 1 | 3 | Johnson, M | female | 27 | 0 | 2 | 347742 | 11.1333 |  | S |
| 11 | 10 | 1 | 2 | Nasser, Mr | female | 14 | 1 | 0 | 237736 | 30.0708 |  | C |
| 12 | 11 | 1 | 3 | Sandstrom | female | 4 | 1 | 1 | PP 9549 | 16.7 | G6 | S |
| 13 | 12 | 1 | 1 | Bonnell, M | female | 58 | 0 | 0 | 113783 | 26.55 | C103 | S |

⋮

# Assignment 2
# Cleaning Titanic Dataset by Pandas

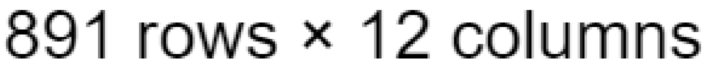
① Problem 1: Load the Titanic dataset from file – Using the Titanic dataset (titanic_rev.csv) that is uploaded on the LMS.

Print the dimension of the dataset.



```
(891, 12)
```

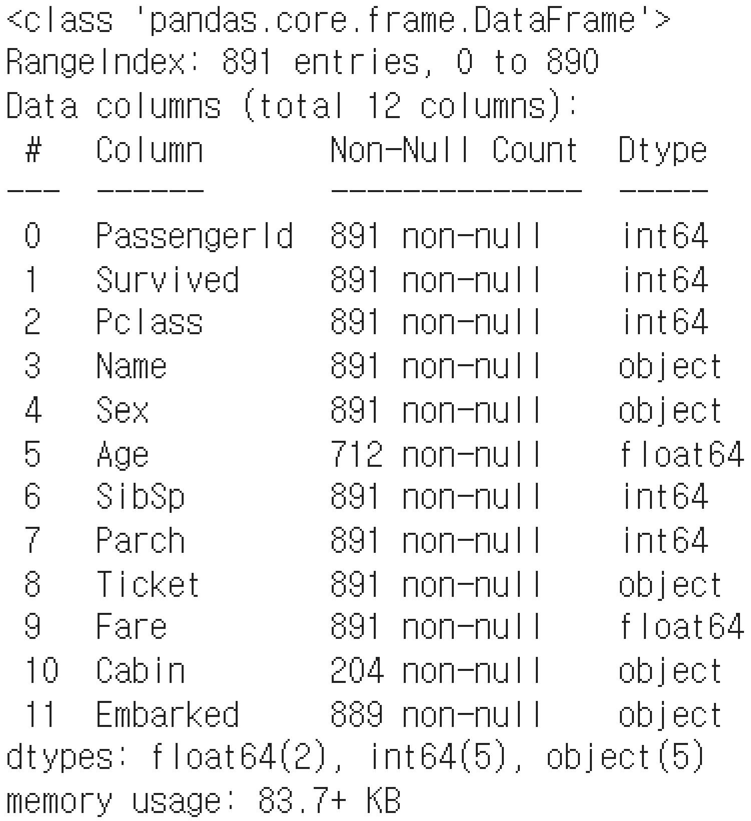
or



```
891 rows × 12 columns
```

# Assignment 2
# Cleaning Titanic Dataset by Pandas

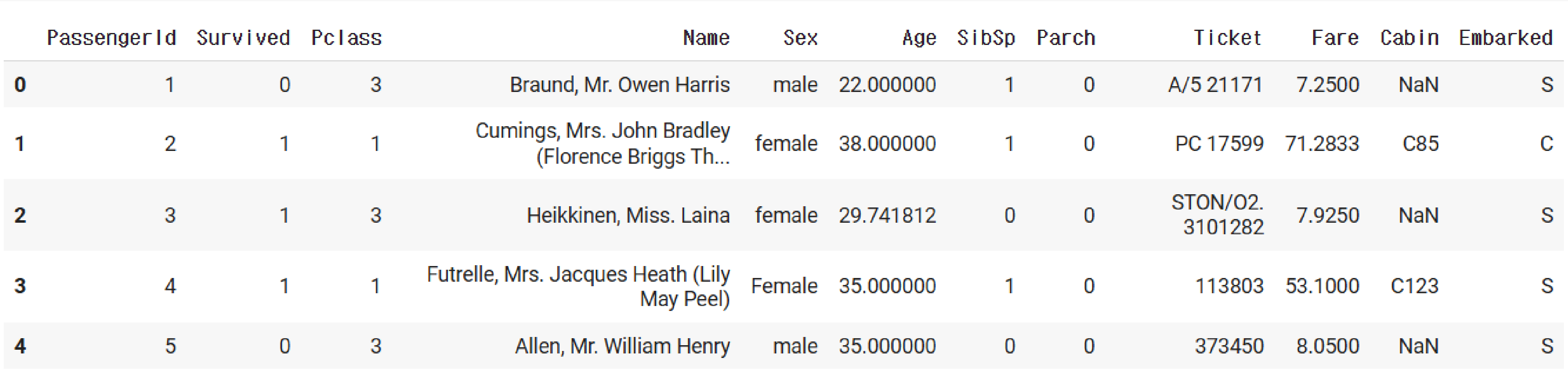
② Problem 2: Print how many non-null values there are in each column



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          712 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

# Assignment 2
# Cleaning Titanic Dataset by Pandas

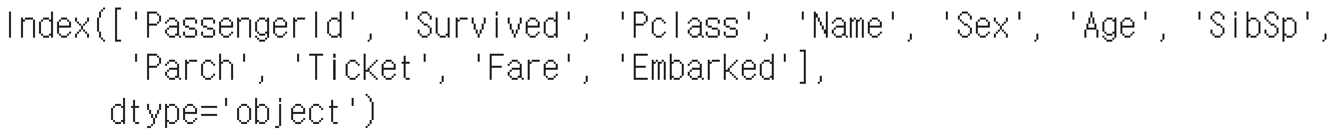
③ Problem 3: Replace the NA value in "Age" column with the *mean* of “Age”. Then, print the *first five* rows.



| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.000000 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.000000 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 29.741812 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | Female | 35.000000 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.000000 | 0 | 0 | 373450 | 8.0500 | NaN | S |

# Assignment 2
# Cleaning Titanic Dataset by Pandas

④ Problem 4: Remove the ‘Cabin’ column. Then, print the column labels. Save the df column with removing Cabin for next problem.

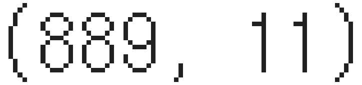


```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Embarked'],
      dtype='object')
```

# Assignment 2

## Cleaning Titanic Dataset by Pandas

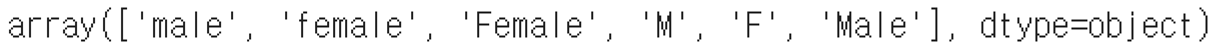
⑤ Problem 5: Remove the rows that have a NA value in the "Embarked" column. Then, print the dimensionality of the DataFrame.



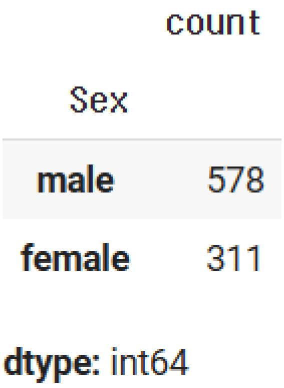
```
(889, 11)
```

# Assignment 2
# Cleaning Titanic Dataset by Pandas

⑥ Problem 6: Print the unique values of 'Sex' Column first. Then, change the value format of the 'Sex' column to use only 'female' or 'male'. Then print the count of unique values in the 'Sex' column.



```
array(['male', 'female', 'Female', 'M', 'F', 'Male'], dtype=object)
```



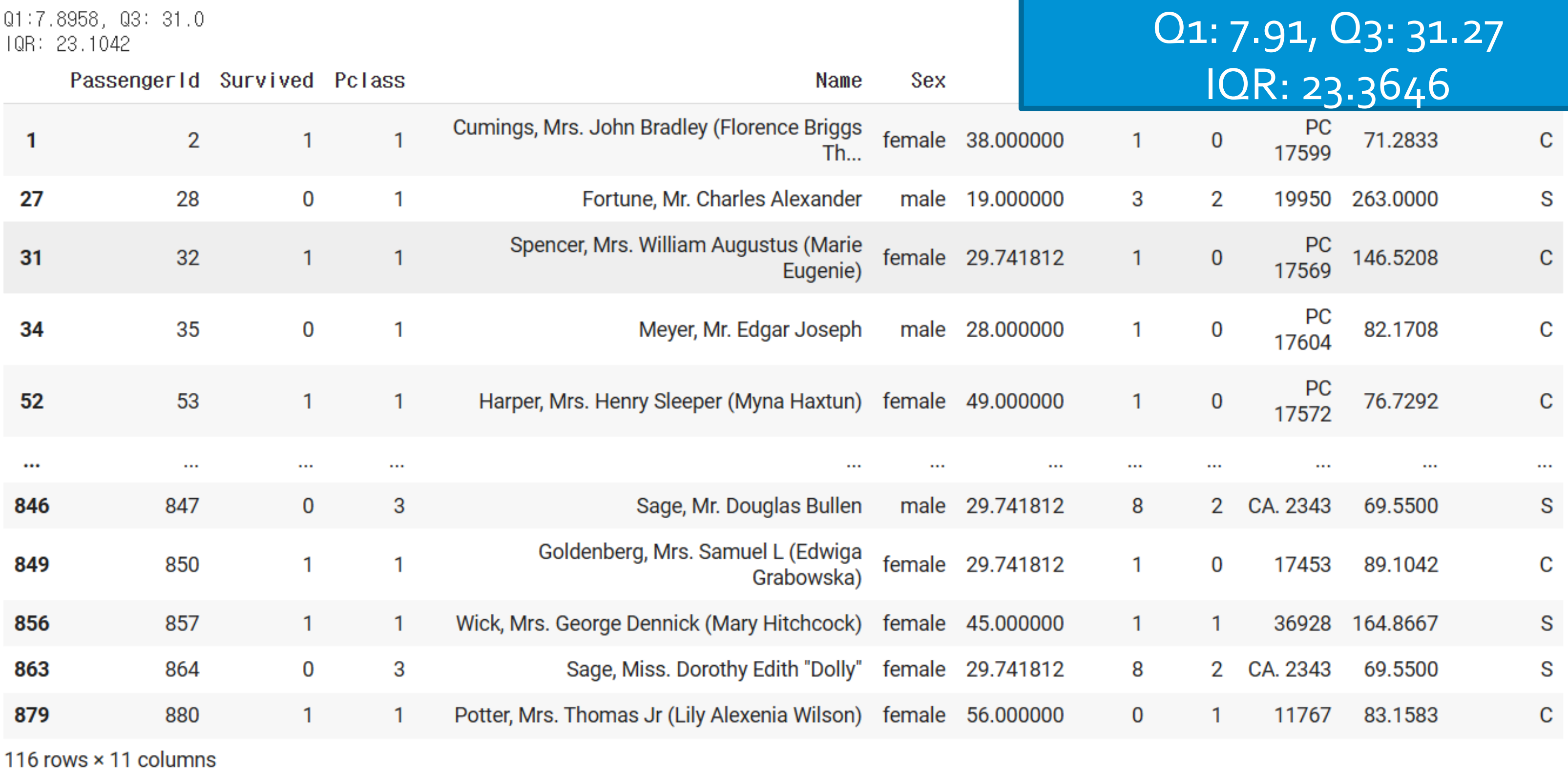
```
          count
   Sex
  male      578
female      311

dtype: int64
```

# Assignment 2
# Cleaning Titanic Dataset by Pandas

⑦ Problem 7: Find outliers in the "Fare" column using the InterQuartile Range (IQR) method. At first print Q1, Q3 and IQR of "Fare" columns. And then print only the rows corresponding to the outliers.

Another answer:
Q1: 7.91, Q3: 31.27
IQR: 23.3646



```
Q1:7.8958, Q3: 31.0
IQR: 23.1042
```

| | PassengerId | Survived | Pclass | Name | Sex | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.000000 | 1 | 0 | PC 17599 | 71.2833 | C |
| **27** | 28 | 0 | 1 | Fortune, Mr. Charles Alexander | male | 19.000000 | 3 | 2 | 19950 | 263.0000 | S |
| **31** | 32 | 1 | 1 | Spencer, Mrs. William Augustus (Marie Eugenie) | female | 29.741812 | 1 | 0 | PC 17569 | 146.5208 | C |
| **34** | 35 | 0 | 1 | Meyer, Mr. Edgar Joseph | male | 28.000000 | 1 | 0 | PC 17604 | 82.1708 | C |
| **52** | 53 | 1 | 1 | Harper, Mrs. Henry Sleeper (Myna Haxtun) | female | 49.000000 | 1 | 0 | PC 17572 | 76.7292 | C |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **846** | 847 | 0 | 3 | Sage, Mr. Douglas Bullen | male | 29.741812 | 8 | 2 | CA. 2343 | 69.5500 | S |
| **849** | 850 | 1 | 1 | Goldenberg, Mrs. Samuel L (Edwiga Grabowska) | female | 29.741812 | 1 | 0 | 17453 | 89.1042 | C |
| **856** | 857 | 1 | 1 | Wick, Mrs. George Dennick (Mary Hitchcock) | female | 45.000000 | 1 | 1 | 36928 | 164.8667 | S |
| **863** | 864 | 0 | 3 | Sage, Miss. Dorothy Edith "Dolly" | female | 29.741812 | 8 | 2 | CA. 2343 | 69.5500 | S |
| **879** | 880 | 1 | 1 | Potter, Mrs. Thomas Jr (Lily Alexenia Wilson) | female | 56.000000 | 0 | 1 | 11767 | 83.1583 | C |

116 rows × 11 columns

# Q & A

김은희 (ehkim@kisti.re.kr)
TA: Yesim Selcuk (yesimselcuk@kisti.re.kr)